



LLM's & Semantic layer

Self-service has entered the chat

Who am I?

- Co-founder / CTO at Zenlytic
- Masters in Data Science from Harvard
- Based in Denver, CO
- Worked in data for 7+ yrs (mostly setting up data stacks)
- Very into rock climbing 🧗





What is a data scientist's job?

...what *isn't* a data scientist's job?

What *is* self-serve?



"Skate to where the puck is going,
not where it has been."

— Wayne Gretzky

2005
SAP BI

View (OLAP)
dashboards

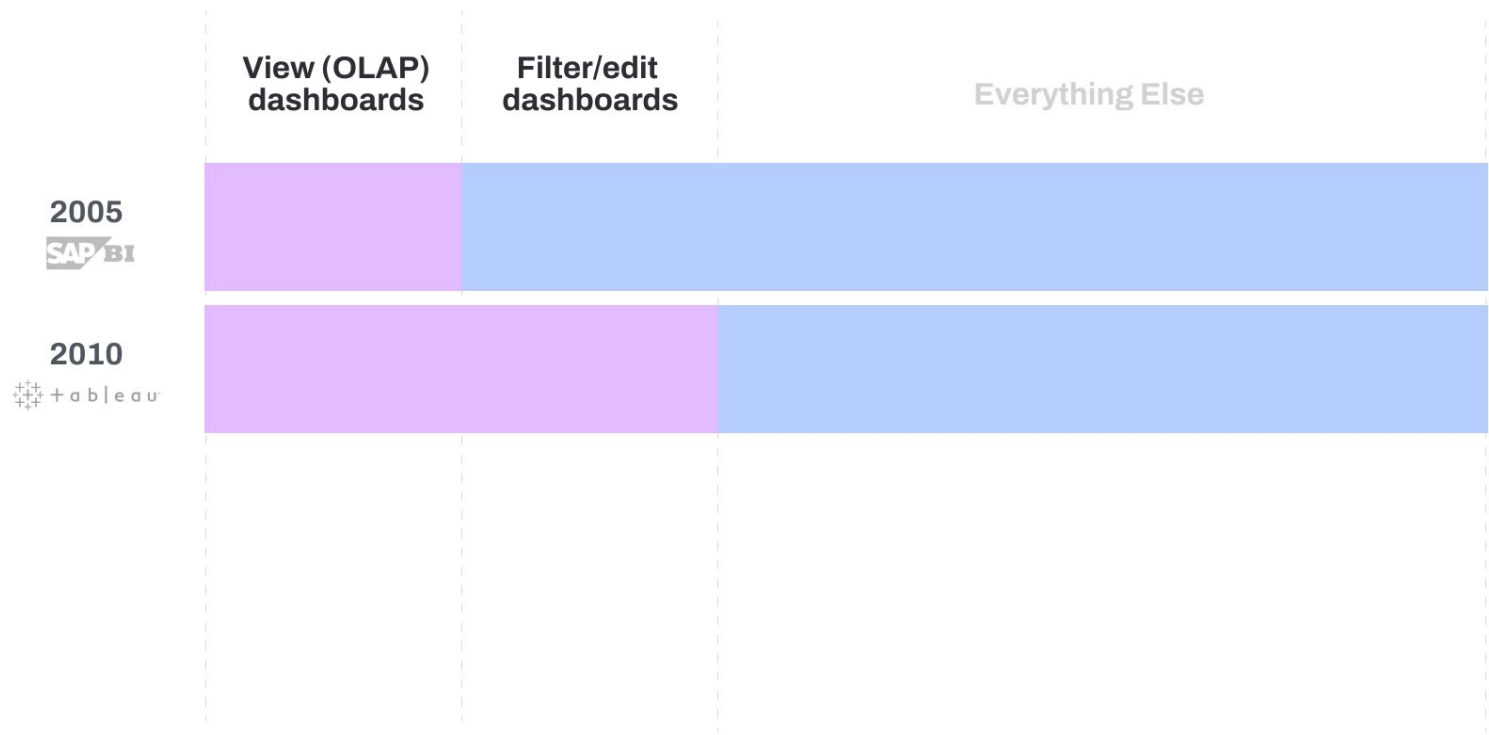
Everything Else



Legend

 Normal people can do it

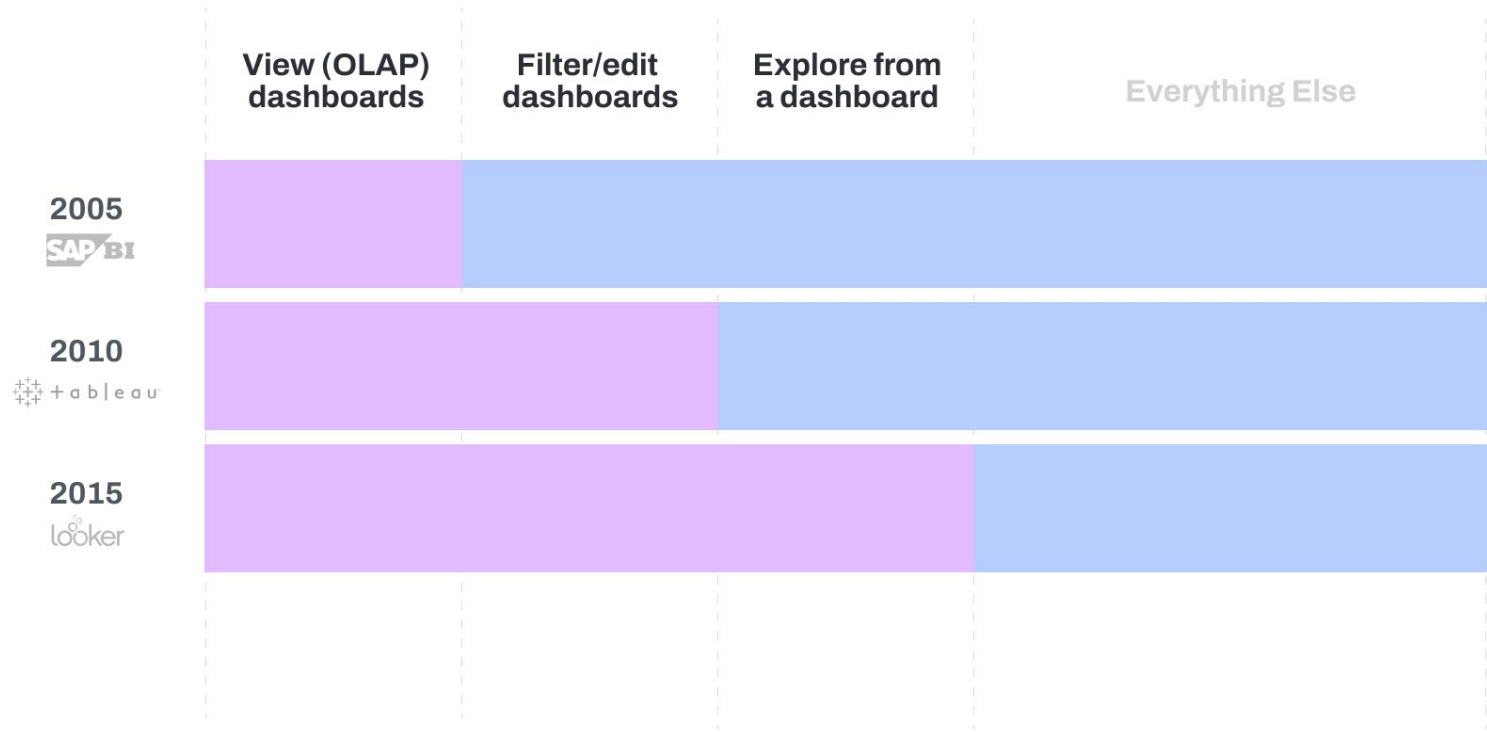
 Data folks do it



Legend

Normal people can do it

Data folks do it



Legend



Normal people can do it



Data folks do it



Legend



Normal people can do it



Data folks do it



Large Language Models

GPT-4





But, GPT isn't good enough



But, GPT isn't good enough

- Do we use `processed_at` or `created_at` for recognizing revenue?



But, GPT isn't good enough

- Do we use processed_at or created_at for recognizing revenue?
- Is net revenue net of refunds or just of discounts?



But, GPT isn't good enough

- Do we use `processed_at` or `created_at` for recognizing revenue?
- Is net revenue net of refunds or just of discounts?
- Why does the customer table fan out when I join on `customer_id`? Isn't that the primary key?



But, GPT isn't good enough

- Do we use `processed_at` or `created_at` for recognizing revenue?
- Is net revenue net of refunds or just of discounts?
- Why does the customer table fan out when I join on `customer_id`? Isn't that the primary key?
- Are 'active' users based on the 'user_status' field or logins in the past X days?



But, GPT isn't good enough

- Do we use processed_at or created_at for recognizing revenue? ✓
- Is net revenue net of refunds or just of discounts? ✓
- Why does the customer table fan out when I join on customer_id? Isn't that the primary key? ✓
- Are 'active' users based on the 'user_status' field or logins in the past X days? ✓

✓: Real world example for me as a human

Even a human can't do these a priori



We can't trust LLM's to pull data

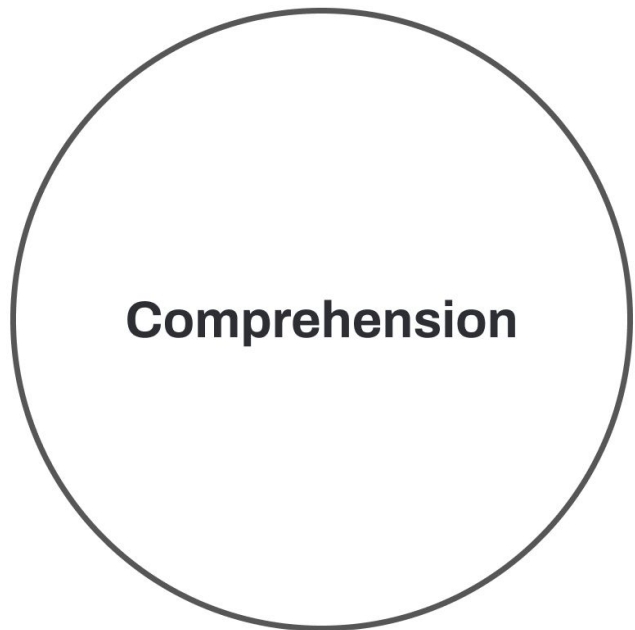
- LLM's generalize
 - Specific choices can't be guaranteed
 - Joins are hard enough for humans
 - Board reporting is at stake here, there are consequences
-
- tl;dr: **text to SQL won't cut it for analytics**



**SELF-SERVICE
ANALYTICS**

LLM

**SEMANTIC
LAYER**



LLMs



Comprehension

LLM's are powerful and impressive, but their core value prop is comprehending the intent of their user and responding

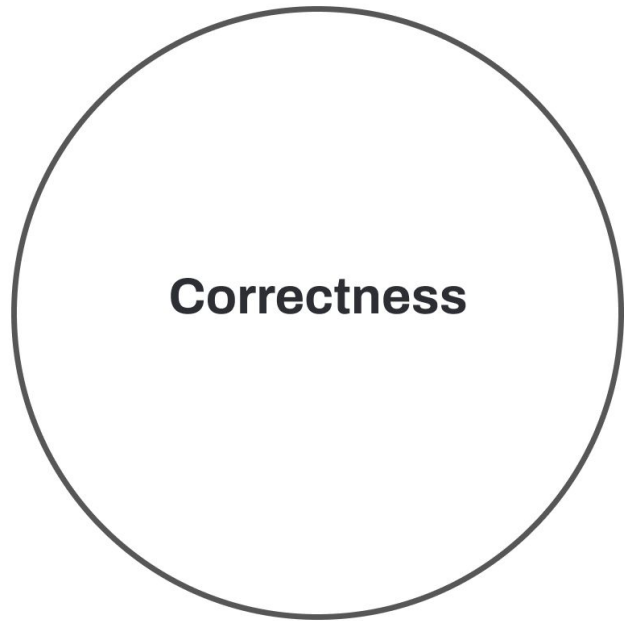


Semantic layer



Correctness

Semantic layers are good for many things, but their core value prop is ensuring queries are correct.



**Semantic
Layer**



Semantic layer is necessary for effective self-serve

Many pre-semantic companies I've worked with aren't even *close* to self-serve.

- They're emailing around GSheets
- They have 9 different definitions of churn
- They're running ad hoc SQL to answer questions about engagement

And their data scientists are working *so hard*



With a semantic layer...

Warby Parker is great at data. But what makes them so good?

Their data team doesn't spend most of their time answering the never ending flow of ad hoc questions.

They build the semantic layer for end users to answer those questions *themselves*



BUT: Semantic layer is not sufficient for self-serve - end-users also need the right UI

Even brands who use semantic layers can't get rid of many ad hoc questions...

- Merging results is hard
- Finding the right Explore is hard
- Iterative question answering is hard

Even with a best-in-class semantic layer, the self-serve UI is too hard for end-users.

Self-serve is a myth

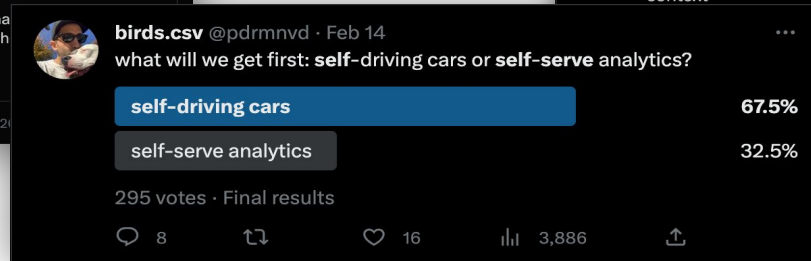
Self-serve has to have both comprehension and correctness to work, and until now we haven't had both



Chris Albon @chriscalbon · Mar 21
Stop trying to make **self-serve analytics** happen, Sarah

Sarah Catanzaro @sarahcat21 · Mar 21
Are LLMs breathing new life into self-service ana we're acutely aware of why self-serve failed in th do better.
[Show this thread](#)

2 1 28 7,324



birds.csv @pdrmnvd · Feb 14
what will we get first: **self-driving cars** or **self-serve analytics**?

self-driving cars	67.5%
self-serve analytics	32.5%

295 votes · Final results

8 16 3,886



Martin Weiss @martinweiss · Mar 22
Data analysts are like: "**Self-serve analytics** are a bad idea. You can't just have anyone crunching numbers."

Also data analysts: "I hate getting so many ad-hoc requests without context"

1 84

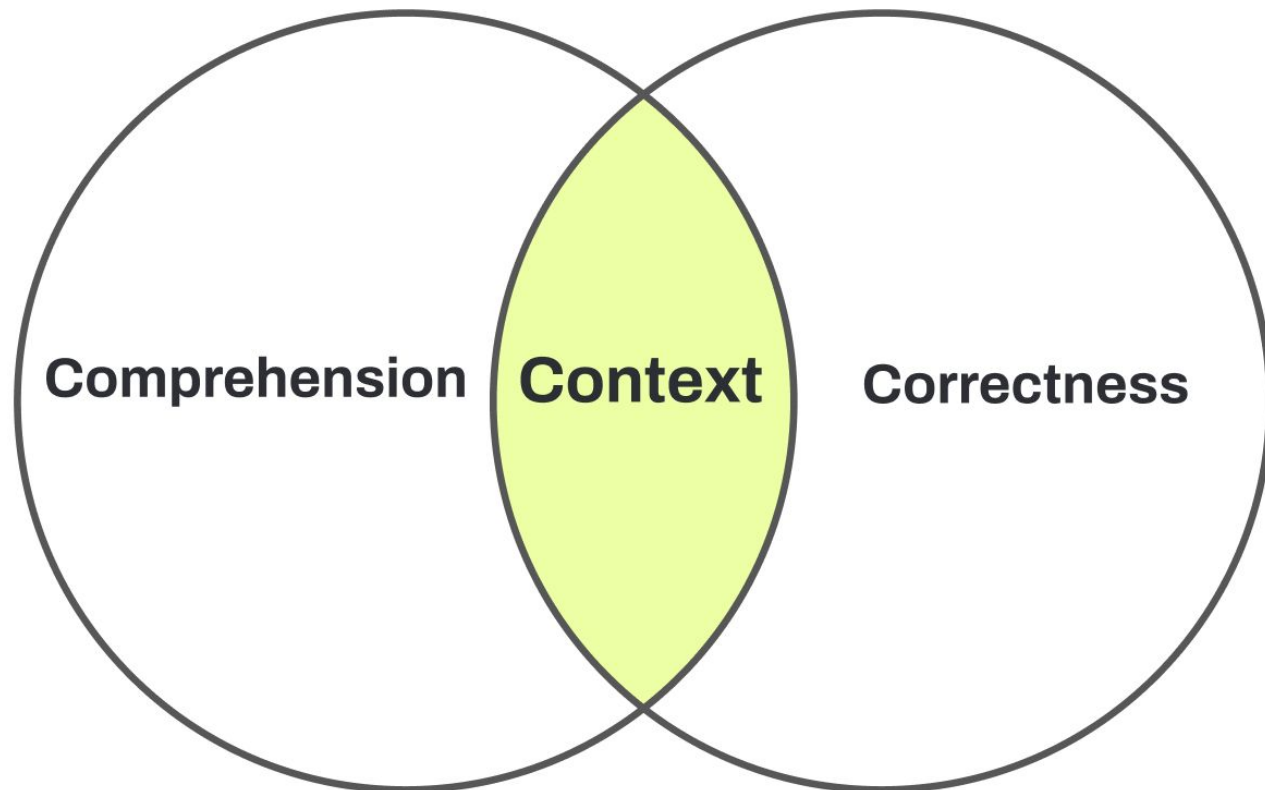


Comprehension

LLMs

Correctness

**Semantic
Layer**



Comprehension

Context

Correctness

LLMs

**Semantic
Layer**

Together, LLM's & Semantic layer have...

Context



Semantic layer *fixes* LLM hallucination

If LLM's are coming up with SQL, hallucination is *catastrophic*.

But if they're using context from the Semantic layer, it's no problem at all

RY

what is 413 divided by 7



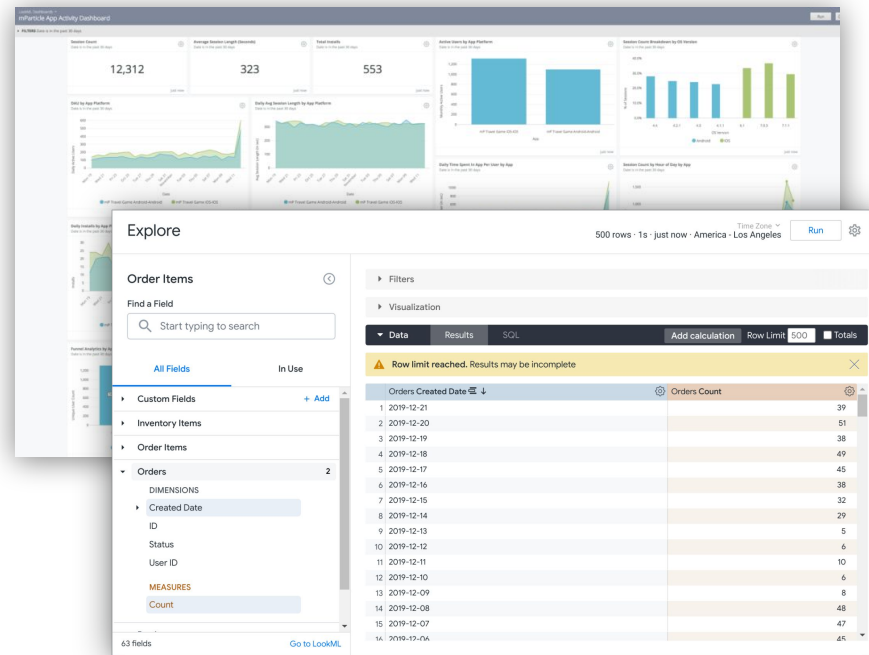
413 divided by 7 is equal to 59 with a remainder of 6.

not quite right...

LLM interface
fixes
Semantic layer
complexity

Even with a Semantic layer, UI's for data are still too hard for end users.

LLM's give that data an interface *everyone* can use



Self serve is like talking to a data scientist

When you combine the correctness of the Semantic layer, with the comprehension of an LLM you can *truly* enable self-serve.



Zenlytic Demo

**What does this mean for
data scientists?**



Legend

Normal people can do it

Data folks do it

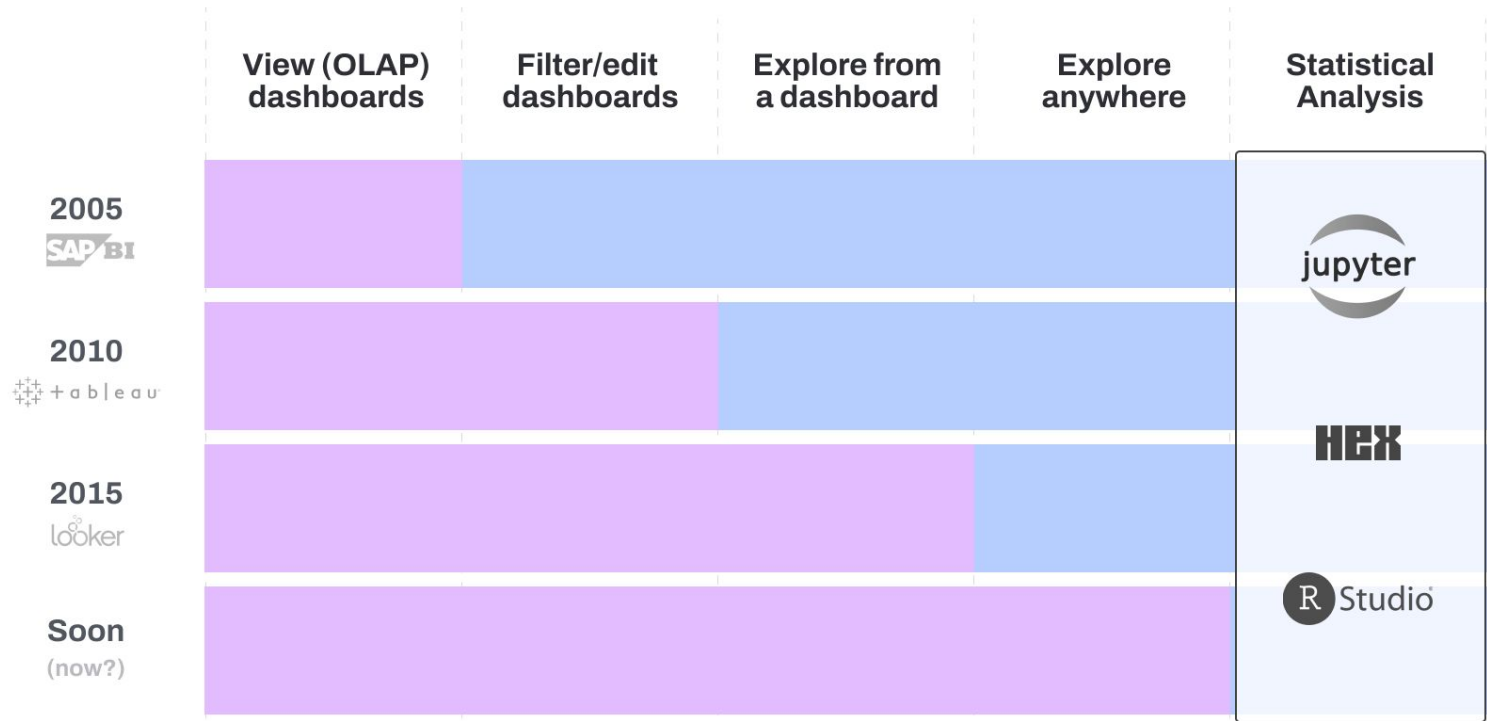


Don't worry, it's good.



Don't worry, it's good.

- More building of these pipes and Semantic layers
- More time spent on the complex things you actually need your education for
- Less time spent on ad hoc question answering



Legend

Normal people can do it

Data folks do it

Questions?





Appendix



For example,

The CEO asks for the number of active users for an investor report

The LLM happily creates a valid query and runs it, based on `status='active'` in the users table

The magic of ✨AI✨

Unfortunately, the status field in the users table doesn't actually mean that they're an active user, it means they activated their account.

'Active' means interaction activity with the app in the last 30 days.

Our CEO shares the great news that active users are up with the board, and then issues a re-statement next month.

Data is hard for humans *and AI*